E/M 99102 WO

Method for the determination of nucleic and/or amino acid sequences

The present invention relates to a method for acquisition of DNA and/or nucleic acid sequences, and in particular a method for acquisition of those DNA and/or nucleic acid sequences of a given species (abbreviated to "species sequences" in the following) which have a potentially increased significance and which thus are research objects which appear particularly worthwhile.

Biosciences, and in particular genetic engineering, have undergone a particularly rapid development in the past years. New processes for producing and duplicating genetic engineering material, such as e.g. the polymerase chain reaction (PCR), and ever better methods for cleavage of genetic material and for identification of the fragments in detail, that is to say the precise sequence of nucleic acids arranged along a gene section, were, for example, the basis for this.

This has meant that the number of gene sections of various species of which the precise structure has been determined has grown ever faster and continues to grow. An aim which is very demanding but will possibly already be achieved in a few years lies in the complete acquisition of the human genome, that is to say acquisition of all the sequences from which human genes are composed, including the precise sequence of nucleic acids within the sequences and the relative arrangement of the individual sequences with respect to one another.

Although the arrangement and positioning of certain sequences can already provide useful additional information in the determination of the functional importance of the sequences in question, mere knowledge of a particular sequence (nucleic acid or DNA sequence) is nevertheless of only very little value if the precise function and importance of the gene section in question has not been recognized and understood. Precisely this, however, plays an ever greater role in scientific research, and in particular in medicine. Thus e.g. certain diseases are very closely linked with the concrete construction of quite particular gene sections, and the precise knowledge of the functional relationship between a particular gene section and the manifestation of a particular clinical picture can therefore be of enormous therapeutic importance, since pharmaceuticals which exactly compensate a pathological deficit can then be developed much more easily. A complete cure may even be brought about, where appropriate, in that by administration of a therapeutic agent, e.g. an inhibitor of a gene product of a gene relevant to a disease, the healthy state

state of equilibrium is achieved again. This of course applies not only to the human species, but in principle to any species of organism, that is to say both to all animal and plant species and to microbiological species.

As already mentioned, the mere discovery of ever new DNA or nucleic acid sequences without knowledge of the functional importance thereof is a relatively useless collection of data, since it is scarcely possible to determine the functional importance of individual sequences or sequence sections in targeted biological or medical research even only approximately at the rate at which new sequences are determined.

Furthermore, in addition to the time required for this, the determination of the functional importance of DNA sequences for which there are no indications at all of the function thereof is also exceptionally cost- and personnel-intensive and therefore ties up many capacities.

On the basis of this prior art, the present invention is based on the object of providing a method for the determination of DNA and/or nucleic acid sequences in which those DNA and/or nucleic acid sequences which have a potentially increased significance are selected out in a targeted manner, that is to say those which can be investigated in a targeted manner in respect of particular functions, in particular in respect of a potential relevance to disease, with considerably less research expenditure than would be possible with the other DNA sequences not selected in this manner.

This object is achieved by the features of claim 1, the dependent claims being advantageous embodiments of the invention by which the selection is refined further and by which additional information which reduces the necessary research expenditure still further is obtained.

The process according to the invention comprises several steps, the sequence of the steps listed below, however, also being at least partly variable. For example, steps b and c could be carried out first, and only then step a of claim 1.

According to step a, in principle any desired species sequences of a species of interest are determined by biological or genetic engineering methods. The species sequences determined are stored in a first databank in a conventional nomenclature as a letter code which consists e.g. of four letters.

According to step b, all known DNA and/or nucleic acid sequences of a given group of biological species or classes are furthermore acquired in a second databank in which in general the functional importances of such sequences are also stored, together with the sequences. Such databanks, which are accessible to the public, additionally sometimes contain further additional information on the individual sequences. Merely for better differentiation, these sequences originating from several species are abbreviated here to "biosequences", while sequences of the species of interest are consistently called "species sequences" here. The given group of species or classes may, but does not have to, contain the species of interest. On the contrary, according to the present invention it is precisely the information on other species contained in such databanks which is selected according to the invention with an ingenious method and which then, by linking with other information sources, with the aid of the method according to the invention indirectly provides indications of the importance of particular sequences of the species in question.

According to step c, the biosequences acquired in a databank according to step b are compared with the species sequences (of the species of interest), which are also already known and possibly stored in the same databank, in a homology test, in which a homology test which is as simple as possible should be used because of the relatively large number of sequences to be compared with one another. If the homology between the known species sequences and the known biosequences then lies above a certain threshold value, according to step d all these biosequences homologous to known species sequences are separated out from the data set to be considered further. The amount of remaining known biosequences has therefore been reduced, compared with the biosequences known in total to the public, not only by a limitation to a group of particular species, but moreover also to those sequences for which no homologous species sequences have yet been determined.

The DNA/nucleic acid sequences stored or newly determined according to step a are then compared in step e with this remaining, reduced set of biosequences in a homology test. The species sequence and the biosequence homologous thereto are expediently matched to one another to confirm the homology and for better understanding of the coinciding sections of the sequences. If the homology lies above a given second limit value, according to step f the biosequences in question, together with at least one linking member which unambiguously identifies the associated biosequence, are stored or issued as a potentially important species sequence.

By linking with one or more particular biosequences of which function descriptions and other additional information are already known, analogous functions of the newly determined species sequences can be searched for in a very targeted manner, and there is also a very high probability of success with relatively low expenditure. This increased probability of success with low expenditure makes the species sequences in question species sequences of potentially increased significance, since other species sequences which are equivalent in structure and length but for which no homologues with known functions exist would require a considerably higher expenditure for determination of their functional importance.

Generally, various information pools are linked with one another by the present invention in a particular strategically favourable manner such that a maximum of information on a sequence is obtained with a minimum of expenditure which is still acceptable in practice. On the other hand, a reciprocal linking, carried out according to conventional mathematical criteria, of all data stored on in each case a sequence and homologues thereof from a relatively large group of biomedical databanks such as are used in the present case would by far exceed all the currently available computer capacities.

In the process according to the invention, not only can successes in the development of medicaments and treatment of diseases therefore be achieved considerably faster and more reliably, this probability of success is increased considerably with a simultaneously reduced research expenditure.

In order to reduce this expenditure still further, in a preferred embodiment of the invention according to a further step g in the databanks accessible to the public references (links) which are stored there to biosequences in the second public databank are acquired, and in particular to the biosequences which have been determined beforehand as homologues to new species sequences, those references which refer to a taxonomically organized databank preferably being evaluated and used. Such a taxonomically organised databank contains keywords to the particular biosequences selected according to standardized scientific criteria, which are then compared according to step h with a given list of keywords, this list in turn being chosen such that it covers the research fields of a user. The biosequence in question and the associated species sequence are thus only obtained in the data set to be defined as worthwhile target objects if agreements exist between a given list of keywords and the keywords in the corresponding databank (third data-

bank) assigned according to taxonomic criteria. The keywords in question, which represent functional importances in a certain manner, then in turn allow more targeted research into the specific properties of a species sequence.

The databank in which newly determined species sequences are stored for further investigation can be a public databank, but probably as a rule a private databank to which in each case only the user or a few users have access, but not the public.

On the other hand, the second databank in which additional information on the biosequences in question and references to other databanks and information stored therein are contained in general has the possibility of access by the public.

A third databank which is particularly suitable for the purposes of the present invention and contains keywords (MeSH terms) selected according to taxonomic criteria is the so-called "MEDLINE" databank. This databank contains on the one hand an identification number for each biomedical literature reference and additional information, together with a number of other data, and inter alia also keywords which are called "medical subject headings". There are moreover links to origins, authors and publications, and so-called RN numbers.

In addition, the MEDLINE databank contains a so-called sequence identifier, which is preferably used as one of the necessary linking members.

It is possible in this manner for a user who originally had only DNA/nucleic acid sequences for which no information at all was known to generate and compile comprehensive information, comprehensive information on a species sequence which characterizes the importance and function of the sequence and allows targeted research being automatically generated by the method according to the invention by the route via homology tests and targeted filtering and separating out of information sources. All species sequences for which functions and importances can be determined in this manner are added to by this additional information. However, they can be taken up again at any time if the data set in the second databank (accessible to the public) has been extended accordingly, so that species sequences initially separated out can also emerge in this manner as worthwhile target objects during a later run.

The homology tests which are carried out between species sequences and biosequences are preferably carried out in a pipeline process, so that complete data sets do not always have to be acquired and managed.

It is furthermore expedient if further databanks, in addition to the databanks already mentioned, are searched for links, in particular with the third databank (MEDLINE), in order also to utilise the additional information from these additional databanks in the event of an appropriate link. These also include, in particular, the databanks called "OMIM" and "KEGG".

Without further statements also, it is assumed that the expert can utilize the above description in the widest scope. Preferred embodiments and examples are therefore to be interpreted merely as a descriptive disclosure which is in no way limiting in any manner.

The complete disclosure of all the Applications, patents and publications listed above and below and the corresponding Application 199 41 606.0, filed on 1st September 1999, are introduced into this Application by reference.

An embodiment example of the invention is explained below with the aid of figures, from which further advantages, features and possible uses of the present invention can be seen. In the figures:

Fig. 1 shows a diagram for reduction of the species sequences determined, such as corresponds to steps a to f in claim 1,

Fig. 2 shows a diagram of databanks and databank links such as are used for further evaluation of information according to the present invention and

Fig. 3 shows a reproduction of a screen with user fields and information fields for a (hypothetical) nucleic acid sequence.

Generally, all newly determined, e.g. in the course of a week, DNA sequences or nucleic acid sequences are initially stored in a databank in a conventional nomenclature (in the standard letter code), an identification number or some other coding for identification of the sequence in question also additionally being assigned and stored at the same time. Further information which is additionally to be co-stored is e.g. the sequence length, the species and other additional informa-

tion directly available together with the determination of such a sequence. The following process steps then proceed automatically. A sequence databank which is accessible to the public and contains DNA and/or nucleic acid sequences of the various species is accessed. By the original input of the species of interest (e.g. *Homo sapiens*), a limitation to a particular group of species of which correlation and functional similarity to gene sections of the species of interest can appropriately be assumed is already implemented here.

The public sequence databank already contains data on the species of interest. A homology test is therefore first [carried out] between the sequences of the species of interest documented in the public databank and the biosequences of the correspondingly selected group of species stored in the same databank. All biosequences which are homologous to the species sequences already stored in the public databank are separated out here, since they were or are evidently already the subject of corresponding research.

The results of this step of the method are expediently recorded, so that when the same operation is repeated, e.g. one week later, all the biosequences which have already been separated out once remain ignored from the beginning, which considerably accelerates the progress of the method. The homology test can then be limited to the newly added biosequences, or conversely the biosequences which have not been separated out beforehand must still be compared with newly added species sequences in a homology test.

The starting data set is thus, however, reduced considerably.

The biosequences still remaining are then compared with the newly determined species sequences in a homology test. As a rule homologous biosequences are found here for some of the newly determined species sequences. A list or table of the species sequences and the newly found homologous biosequences for these is then prepared, and additional information from the public databank, such as e.g. a Medline identity number possibly stored for a known biosequence, is also included in this table or list.

A further step (h) of the method consists of classification of the species sequences issued or stored in step f), i.e. assignment (sorting) into particular classes of sequences by linguistic analysis of text definitions of the additional information stored on the homologous biosequences. This

allows division into partial data sets, in turn only a part of the other databases being appropriate for adding to these.

According to step i, the characteristic information of the particular homologous biosequences which is to be assigned to the potentially important species sequences is added to by acquisition of references (links) relating to the biosequences in the second databank acquired according to step f) to at least one third databank and acquisition of the information on the biosequences mentioned which is stored in the third databank.

The third databank should provide a classification which is taxonomically organized at least into part regions, and is preferably here the so-called MEDLINE databank.

According to the invention, the keywords assigned to the particular biosequences according to taxonomic criteria are compared with a given list or file of keywords and coinciding keywords as well as the biosequences in question and the homologous species sequences or in each case an identification thereof for which keywords which coincide with the given list of keywords have been found are issued.

In addition to the MEDLINE databank or also as a substitute for this, information from further databanks which are chosen e.g. from the group consisting of the Unigene, Genemap and GDB (new) as well as OMIM, KEGG and UMLS databanks is also used.

The species of interest is primarily that of *Homo sapiens*, but the method according to the invention can equally be used with a substantially similar aim for another species.

The course and the result of a hypothetical embodiment example will now be explained in somewhat more detail with reference to the figures. As already mentioned, according to step c in patent claim 1, species sequences of the species of interest which are already known are compared in a homology test with the biosequences which belong to a given group of biosequences and are stored in the second databank. This step is called "blastx humprot" in fig. 1. If homologous sequences have been found, a particular status (in this case status = 2) is assigned to the biosequences homologous to the species sequences already known, and these biosequences are identified accordingly and separated out of the pool of interest of the second databank.

With the species sequences which have been determined according to step a, a further homology test is then carried out with the biosequences remaining from the second databank which until then have not yet been determined as homologues to the known species sequences. This step is called "blastn proprietary genes" in fig. 1. If homologous biosequences have been found, the best possible match and alignment is carried out (this step is called "bestfit" in fig. 1) and the data which characterize the match, length and alignment are stored, together with the sequence in question. The status 0 assigned to the corresponding biosequences means that these biosequences continue to remain in the data pool of interest.

Likewise, those biosequences in the reduced data pool of interest for which homologues were to be found neither among the species sequences determined nor among the species sequences already known also remain.

Data sets to which homologous biosequences corresponding to newly determined species sequences are assigned are generated in this manner. The user of the system according to the invention expediently uses this from a screen workplace with appropriate equipment. A screen display which shows a hypothetical result of a determination of potentially important species sequences according to the invention is shown in diagram form in fig. 3. It is to be pointed out here, however, that the result shown is not a real result, but merely a hypothetical artificially synthesized result from which, however, in principle all the essential steps and results of a typical embodiment example can be seen.

The screen shows at the left edge a series of command and parameter fields which the user can use. For example, in field 1.2 he selects a limit value parameter which indicates the minimum length of homology between the species sequence and biosequence which coincides with the nucleic acids of the homologous sequence according to the homology test and best possible fit. In field 1.3 the limit value of a percentage agreement is shown. A keyword which is to be searched for in connection with the corresponding homologous sequences can be entered e.g. in field 1.4.

The other user fields are self-explanatory.

When the user has chosen appropriate parameters and the basic program starts, after a short time he/she obtains a list of species sequences which have one or more biosequence homologues

which meet the criteria of the user's input. For example, fig. 3 shows that 124 species sequences have one or more biosequences which are homologous with a percentage identical nature of greater than 95% and have a homology length of greater than 500 base pairs. Moreover, the entries have MeSH terms which are chiefly associated with the CNS (central nervous system). Of the 124 entries, fig. 3 shows the fifth species sequence, which is designated the number sequence 44567. The biosequences which are homologous with the species sequence are shown in the right screen half under "seeds". Several steps are necessary here in order to be able to generate this assignment of individual data from extensive files to a particular given species sequence, including the large amount of additional information, but these proceed automatically in an appropriate program, the progress being explained in diagram form on fig. 2. From the homology test called "blast proprietary genes" in fig. 1 and from the resulting homologues in the second databank, so-called Genbank identifiers (Genbank ID) can be determined from the second databank and are in turn also filed in other databanks, so that a relationship is established between various nucleic and/or amino acid sequences and other information stored in the databanks.

The Medline databank and the MEDLINE identifier ("Medline ID" block) recorded in many other databanks have a key function here. The sequences given under "seeds" are characterized by a Genbank identifier. These entries named by the Genbank identifier can also contain, inter alia, Medline identifiers. The title of the corresponding entries can be determined from the MEDLINE databank with the aid of this Medline identifier. Furthermore, references to particular enzymes connected with the gene section in question are often also filed in this databank, and the biochemical reaction pathways influenced by these enzymes in turn result from these. Further information from other databanks, e.g. pathological information, the location of genes on particular chromosome sections etc., can moreover be obtained via the MEDLINE identifier.

After running an appropriate program, a whole series of information which, in addition to the probable location of the newly determined species sequence, gives a whole range of references to its function, organ distribution and relevance to disease, is then reproduced on the screen. In the present case, which, as already mentioned, shows only hypothetical information on a species sequence, alongside the sequence 44567, for example, the biochemical name and the date of issue of the information and in 17q23 the position of the gene section on a chromosome can be seen. Genes located on the same chromosome arm are shown underneath. Information on clusters of gene fragments (EST clusters), which are identified by a particular number (Hs.198237), originates from the UNIGENE databank. The number of ESTs in this cluster in relation to the

total number of components of the present sequence is given as 54/82. Proangiotensin-angiotensin indicates the most probable metabolic pathways or chemical reactions to which the known biosequences belong. BRAIN furthermore indicates that organ in which the sequences in question are found most frequently. The organ distribution of the EST components is illustrated by different bar lengths. The most probable area of a disease indication which has been determined in connection with the data comparison is given as CNS. In the left half a horizontal row of bars can also be seen, the length of these bars in each case indicating agreements between the species sequence and the associated biosequences or sequence sections shown in the corresponding lines. In addition, the biosequences are listed individually under "seeds", including their percentage agreement and the length of the coinciding sequence sections. The titles of appropriate journals, the enzymes and various keywords are also given.

In the present example, information obtained by the linking according to the invention via various identifiers, keyword searches and taxonomic evaluation of databanks was determined from most of the databanks shown in fig. 2, with the exception of the blocks designated UMLS, SNOMED and ICD9-CM. The Knowledge Interchange Format (KIF) is used to store the information obtained from the method. This format can be used by various knowledge engineering tools, such as e.g. Ontolingua, in order to generate, inter alia, HTML or XML files and to use more extensive methods of artificial intelligence (AI).